Keywords: networks, gene regulation, signaling logic, pathways, alignment, network comparison.

Functional alignment of regulatory networks: A study of temperate phages

Ala Trusina,[1, 2] Kim Sneppen,[3, *] Ian B. Dodd,[4] Keith E. Shearwin,[4] and J. Barry Egan[4]

[1]*Department of Theoretical Physics, Umeå University, 901 87 Umeå, Sweden*
[2]*NORDITA, Blegdamsvej 17, Dk 2100, Copenhagen, Denmark*[†]
[3]*NBI, Blegdamsvej 17, Dk 2100, Copenhagen, Denmark*
[4]*Discipline of Biochemistry, School of Molecular and Biomedical Science,*
University of Adelaide, South Australia 5005, Australia

(Dated: February 9, 2008)

The relationship between the design and functionality of molecular networks is now a key issue in biology. Comparison of regulatory networks performing similar tasks can give insights into how network architecture is constrained by the functions it directs. We here discuss methods of network comparison based on network architecture and signaling logic. Introducing local and global signaling scores for the difference between two networks we quantify similarities between evolutionary closely and distantly related bacteriophages. Despite the large evolutionary separation between phage λ and 186 their networks are found to be similar when difference is measured in terms of global signaling. We finally discuss how network alignment can be used to to pinpoint protein similarities viewed from the network perspective.

SYNOPSIS

Networks of interacting genes and proteins orchestrate the complex functions of every living cell. Decoding the logic of these biochemical circuits is a central challenge facing biology today. Trusina et al. describe a mathematical method for aligning two regulatory networks based on their signaling properties, and apply it to a case study of three bacteriophages, simple biological "computers" whose genetics are exceptionally well characterized. The comparison reveals a surprising similarity between regulatory networks of the creatures, even when they have very distant evolutionary relationships. The method introduced here should be applicable to other networks, and thus help illuminate the computational substructures of living systems.

INTRODUCTION

The functioning of living organisms is based on an intricate network of genes and proteins regulating each other. Various organisms differ due not only to differences in the constituting components (genes/proteins), but also because of the organization of these regulatory networks. It is therefore important to address similarities and differences not only in protein sequences but also in the interaction patterns of the proteins. Thus, large scale analysis of protein-protein and protein-DNA interactions have provided insight into the local design features of subcellular signaling [1, 2, 3]; network alignment based on sequence similarities permits alignment of related motifs [4, 5].

Here we suggest to compare networks through an alignment method that is based solely on network architecture and signaling logic, and thus does not rely on sequence similarity of the involved proteins.

As a study case we consider the regulatory networks of two very well-characterized temperate bacteriophages of *E. coli*, λ and 186 (Fig. 1). These two phages represent two distinct classes of temperate bacteriophages: the lambdoid phages - which include λ, P22, 434, HK97 and HK022, and the P2 group - which includes P2, 186, HP1, K139 and PSP3. λ and 186, are not detectably related in sequence and have different genome organizations. Using tBLASTx [6] to compare all of the reading frames, there are only two clearly homologous protein pairs - the λ endolysin R (P03706)/186 (PO80309) (E-score = 10^{-34}) and a pair of early lytic proteins of unknown function (E-score = 2×10^{-4}). No significant similarity was detectable at the nucleotide level (using BLASTn, [6]). On the genome level, the arrangement of genes, promoters and operators is very different [7, 8, 9, 10]. As a control of methodology, we also consider the P22 phage, which as a member of the lambdoid family allows us to compare topologies of evolutionarily related networks.

As temperate phage, both 186 and lambda can be in two states: a lytic state where many proteins are active in the replication of the phage DNA and the construction and release of virus particles; and a lysogenic state where the phage genome is integrated into the bacterial chromosome and only a few proteins are active. For both phages, three core proteins (CI (P03034), Cro (P03040) and CII (P03042) in λ, and CI (P08707), Apl (P21681) and CII (P21678) in 186) do the main computations, with the switch into lysogeny being coordinated by CII and the reverse switch into the lytic mode initiated by activation of the host SOS response protein RecA (P03017). The gene regulatory networks of all temperate phage have evolved to provide lysogenic and lytic states, and more than that, to switch from one state to another when particular signals have been received from bacterial proteins,

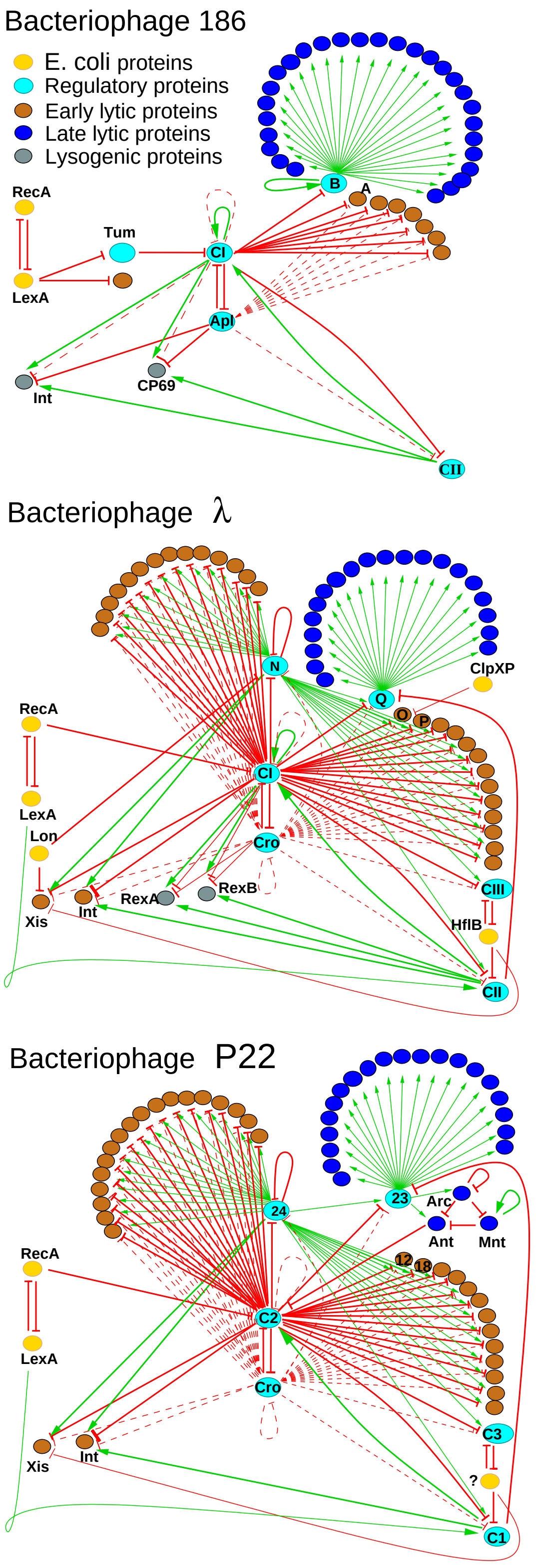

FIG. 1: The genetic regulatory networks for phage 186 for phage λ and P22 all of which are temperate and infect *Escherichia coli*. The proteins are colored according to their functions and expression mode in the lysis-lysogeny life cycle of the phages. We summarize the influence of one protein on another by either a green (positive, e.g., transcriptional activation) or a red (negative, e.g., repression) arrow. The dashed lines show relatively weak regulations. Database entry for λ genome is J02459, for 186 genome – U32222, and for P22 genome it is NC_002371.

and thus effectively perform the same function.

Given that 186 and λ are both temperate, i.e. performing similar function, but are evolutionary separated, we asked whether we can detect structural similarities and what is the scale at which these similarities are detectable ?

RESULTS

Visual comparison of the 186 and λ networks (Fig. 1), suggests both strong similarities but also major differences. One way to quantify the similarity of two networks is by edit distance [11]. Assume that we know which nodes (here, proteins) in network A and B should be paired. For networks of the same size, we define edit distance as the number of insertions or removals of edges (regulatory connections) one has to perform on network A to obtain B. This is quantified through

$$D_E(A,B) = \sum_{i,j} |A_{ij} - B_{ij}| \,, \qquad (1)$$

The elements A_{ij} and B_{ij} specify whether the direct regulation of i on protein j is positive, negative or absent and are constructed such that each element can keep both positive and negative links (for details see eq. (2) below).

In case we do not know which nodes in networks A and B should be paired, we find the optimal identification by minimizing D_E as described in *Materials and Methods* section. This yields the minimal distance between the networks, as well as an optimal alignment of the individual nodes. This distance we call the edit difference.

The minimal edit difference between *related* phages is small $D_E(\lambda, P22) = 18$, compared to the larger scores for evolutionary separated phages, see table I. The $D_E = 18$ means that, the λ network of 62 proteins and 144 connections can be constructed by making 18 edits of the connections in a 62 protein subset of the 67 protein $P22$ network (adding/removing a link is a single edit, changing the sign of a connections needs two edits). To get an idea of the significance of the obtained D_E values, we compare with optimal alignments of 500 randomized versions of the two networks. The randomization procedure was designed to conserve the local properties of the networks in order to try to keep their general biological features. Firstly, the core-hub topology common in biological networks [3] was maintained by conserving for each protein the number of its regulators (inputs) and the number of proteins regulated by it (outputs). Secondly, the number of each sign (positive and negative) of the input and output connections was kept for each node.

The constrain of preserving the local properties does not fix the network completely: while keeping the number of positively/negatively regulated proteins one can still change which exactly of them are being regulated.

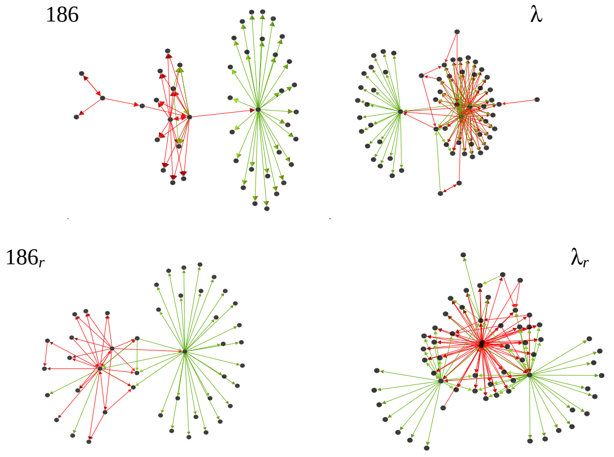

FIG. 2: Illustration of the differences between the real 186 and λ networks (top) and an example of their randomized counterparts (bottom). These examples of randomized networks show that it possible to preserve local properties, yet obtain different network structures.

The structure of the resulting random networks is rather different, as seen in the examples shown in Figure 2.

Overall we find that D_E scores between any pair of randomized networks are similar. When comparing scores between real network, with that of their random counterparts in table I one see no clear trend. In particular the differences between these randomized versions, λ_r and 186_r were indistinguishable from that of the real networks: $D_E(186_r, \lambda_r) = 32 \pm 2$.

We reasoned that the functional similarity of networks might be better reflected in a less local measure of functionality. We therefore introduce a signaling difference D_S, which aims at capturing both direct (as in D_E) and also indirect regulation through a sequence of intermediate proteins. For each pair of proteins (i, j), we consider whether i sends a signal to j, and if so whether the signal along the shortest path is positive or negative. In this spirit we define the sign of a signal as the product of the signs of all links on the shortest path from i to j. An example where this procedure nicely reflects the functionality in terms of its “Boolean” logic [15] is found in the pathway from RecA to CI in the two phages . In λ, active RecA directly catalyzes self cleavage of CI [12]; whereas in 186, RecA acts through the degradation of LexA (P03033), that in turn represses the protein Tum [9] (P41063), which in the absence of repression binds CI and prevents it from performing its function. Thus the simple -1 signal in λ is in 186 replaced by a signaling consisting of $(-1) \times (-1) \times (-1) = -1$. In other words repressing a repressor is effectively an activation.

Because the regulation of one protein by another may be positive through one series of links and negative through another, two matrices were used for each network, one for positive signals (A^{S+} and B^{S+}) and one for negative signals (A^{S-} and B^{S-}). If the effect of protein i on protein j is only positive, then 1 is placed into A^{S+}_{ij} and 0 into A^{S-}_{ij}. If the effect is only negative, then 0 is placed into A^{S+}_{ij} and 1 into A^{S-}_{ij}. If there are positive and negative signals along paths of equal length (e.g. from RecA to λ CII via LexA or CI), then 1 is placed into both matrices. Observe that when positive and negative signals come to the same node, they are not canceling each other. This is intended, as often signals will arrive at different times or at different conditions [16].

netA, netB	D_E	P_E	D_S	P_S
λ, 186	33		43	
λ_r, 186_r	32 ± 2	0.27	109 ± 33	0.01
λ, $P22$	18		106	
λ_r,$P22_r$	33 ± 4	0.00	255 ± 55	0.00
$P22$, 186	25		97	
$P22_r$, 186_r	31 ± 1	0.00	161 ± 36	0.03

TABLE I: The overall difference measures D_E, D_S between the networks, with respective P-scores as defined in text.

The signaling difference between two networks A and B is then defined as

$$D_S(A, B) \;=\; \sum_{ij} |A^{S+}_{ij} - B^{S+}_{ij}| \;+\; |A^{S-}_{ij} - B^{S-}_{ij}| \quad (2)$$

which takes into account differences in both positive and negative signaling along the shortest paths between any pair of nodes. Like D_E, the minimum difference D_S is calculated by optimizing which proteins in 186 should be identified with which proteins in λ, and in addition which λ proteins should be excluded. Excluding a protein means that the signaling to and from that protein is not counted in D_S, whereas signaling across the excluded protein is included.

Optimizing protein alignment based on signaling, we find $D_S(186, \lambda) = 43$. Again, the significance of this difference was determined by repeatedly performing randomization of the networks as described above, creating the A^{S+} and A^{S-} matrices and obtaining the minimal D_S. The differences between random networks, $D_S(186_r, \lambda_r) = 109 \pm 33$, is much larger than between the real networks. This is further quantified by a P-score, $P(D_S > D_S(random)) = 0.01$, defined as the probability that two randomized networks will have a smaller difference than that between the real networks.

Thus all three networks are similar in their signaling pattern. To confirm that this signaling similarity is not generally conserved among biological networks, we have compared the phage networks with other networks that perform different functions (e.g. the *S. cerevisiae*, [17], cell cycle network and the *B. subtilis* competence network, [18]). We found that D_S is much larger and the P-scores are close to 1 in these alignments, indicating that the low signaling difference between the phage networks is a special property of these functionally similar networks.

We have also considered other variants of the difference measures, in particular including all non-repetitive paths between pairs of proteins, with all paths weighted equally. In that case we also find that $D_{S-all}(\lambda, 186) = 390$ between real network is smaller than $D_{S-all}(\lambda_r, 186_r) = 583 \pm 122$ between the randomized counterparts. Also, using the shortest paths, we have investigated differences between networks where weak links (the dashed ones in Fig. 1) are weighted less (by a factor 0.5 or removed altogether). D_S scores between networks got smaller, but overall significance remained similar.

DISCUSSION

The pathway related D_S score allowed us to identify significant similarity between two very distantly related biological networks , see table I. In contrast, the edit difference measure, which looks only at the local wiring structure, is sometimes blind to this more global "homology". Thus although edit difference partially captures network similarities through a patchwork of local matchings, it is less sensitive to pathway disruptions.

It is not clear whether the functional similarity between the lambda and 186 networks detected by the D_S measure is a result of convergent evolution or is a remnant of a shared ancestral network. Under either scenario it is clear that the two network structures must be strongly constrained by functional requirements, given the evolutionary separation of the two phages. A potential bias should be noted here: knowledge of the three phage networks is not complete, even for λ, and it is thus possible that some of the observed similarity in the networks is due to knowledge of connections in one phage network having influenced the discovery of connections in the others.

The D_S alignment allows us to address the role of various proteins in pathway disruptions. Figure 3 line up the λ and 186 proteins on the basis of pre-existing knowledge of their function or mode of expression and have indicated the optimal D_S alignment and the contribution of each pair to the signaling difference. The two alignments show good matches for late lytic genes as well as for the regulators CI, CII and B from 186 aligned with CI, CII and Q in λ. Thus in general functions of proteins in one network teaches us about protein properties in the other network. The lack of a good match between Apl (in 186) and Cro (in λ), is due to the weak links from Cro, and reflects a different functional role of Cro and Apl in the late lytic development of phages. Insisting on alignment of Cro with Apl results in $D_S = 219$, thus emphasizing the particular role of Cro as a repressor of late lysis in λ.

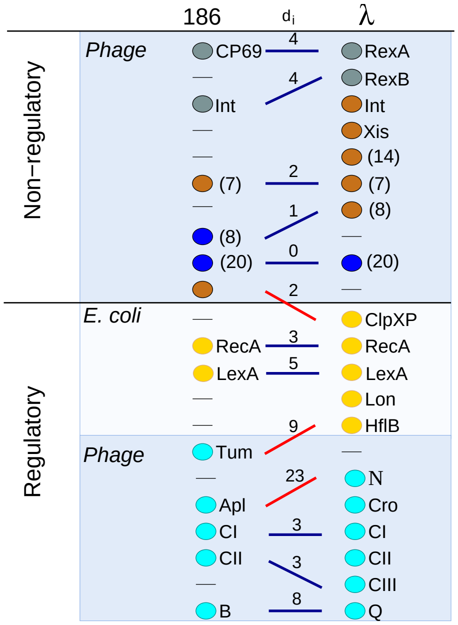

FIG. 3: Alignment of two phage networks. Placement of proteins is based on our knowledge [7, 8, 9, 12, 13] and the lines connecting them are associated to the minimal D_S alignment. Proteins that perform similar functions or are regulated similarly are placed on the same level; thus horizontal lines mark ideal matching. Blue lines correspond to meaningful alignments, red lines are the misalignments. The numbers above the lines, d_i, reflect the differences in signaling between the aligned proteins and are the contributions to the minimal difference $D_S = \frac{1}{2}\sum_i d_i = 43$. The numbers in the parentheses indicate multiple equivalent proteins, making the sum of all shown signaling differences equal to $2 \cdot 43$. The key regulators RecA, LexA and CI are identified correctly whereas the misidentification of CII with CIII is reasonable since both favor entry into lysogeny through the same pathway. The major discrepancy is associated to different roles of Cro and Apl during lysis (the weak links from Cro to Q and N in λ).

Comparison of molecular networks is becoming an important element of modern systems biology, both with regards to predictions of eventual missing links [19], and for increasing our understanding of functionality of information processing in the networks. The here presented alignment methods address the similarities on a local, respectively larger scale, associated to signaling across networks.

In this regards we found that evolutionary relationships ($\lambda - P22$) imply similar local regulation, with low D_E score. For all temperate phages, evolved to do similar "computation", their regulatory networks are found to be similar when viewed from a more global perspective where both direct and indirect signals are included (low D_S score compared to random expectation). Thus the mechanistic and structural differences on the scale of genome and promoter organization disappear when considering the large scale of the protein regulatory networks. Going beyond immediate regulations allows to capture functional similarity in the most robust way.

MATERIALS AND METHODS

The present papers is based on the data on three bacteriophages λ (accession no. J02459), P22 (NC_002371) and 186(U32222). The regulatory networks were compiled from these database entries and various literature sources: λ ([7, 13, 20, 21] and references therein), for 186 ([8, 9, 10] and references therein), for P22 ([14] and references therein).

In the *Results* section we define two differences scores, D_E and D_S between a pair of networks A and B. Provided that we know which proteins in A should be identified with which in B, the scores are calculated as in Eq. 1 respectively Eq. 2. In case we do not know which nodes in networks A and B should be paired, we need to find the optimal identification of nodes between them. To do so, we define an alignment procedure through the Metropolis algorithm [22] designed to reach the minimal distance D between the networks: Given two nodes and their corresponding partners in the other network the elementary step is to switch partners and reevaluate the distance. Iterating this procedure and using simulated annealing [23] the method converges to a global minimum.

If the two networks are of different size we count only the contribution from a number of nodes given by the smaller of the two networks. In the larger network these nodes are selected to minimize the distance using the above algorithm.

We would like to note that the above method is not intended to reflect any evolutionary process, but is used to find the optimal mapping of pairs of proteins that look similar from the network perspective. The method is limited by the network size, and in practice works for networks below 200 nodes.

The realization of the alignment algorithm in form of the Java applet is available at http://www.cmol.nbi.dk/models/compar/compar.html.

ACKNOWLEDGMENT

We warmly thank S. Brown, S. Krishna and S. Strogatz for constructive comments on the manuscript. The work was supported by the Swed. Res. Council Grants No. 621 2003 6290, 629 2002 6258 and “Danmarks Grundforskningsfond” through the center “Models of Life” at the Niels Bohr Institute.
Work in the Egan lab is supported by the US NIH.

[*] Electronic address: sneppen@nbi.dk

[†] URL: `cmol.nbi.dk`

[1] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabsi A.-L. (2000) The large-scale organization of metabolic networks. *Nature* 407, 651-654.

[2] Shen-Orr, S., Milo, R., Mangan, S., and Alon U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* 31, 64-68.

[3] Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* 296, 910-913.

[4] Kelley, P.B., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., and Ideker, T. (2004). PathBLAST: a tool for alignment of protein interaction networks *Nucleic Acids Research* 32, 83-88 (2004).

[5] Berg, J., and Lassig, M. (2004). Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. U S A* 101, 14689-14694.

[6] Tatusova, T.A., and Maddenm, T.L. (1999). Blast 2 sequences - a new tool for comparing protein and nucleotide sequences, *FEMS Microbiol Lett.* 174, 247-250.

[7] Dodd, I.B., Shearwin K.E., and Egan J.B. Revisited gene regulation in bacteriophage lambda. *Curr Opin Genet Dev.* 2005 **15** (2):145-52.

[8] Portelli, R., Dodd, I.B., Xue, Q., and Egan, J.B. (1998). The late-expressed region of the temperate coliphage 186 genome. *Virology*, 248, 117-130.

[9] Shearwin, K.E., Brumby, A.M., Egan, J.B. (1998) The Tum protein of coliphage 186 is an antirepressor. *J Biol Chem.* 273, 5708-15.

[10] Dodd, I.B., Egan, J.B. (2002). Action at a distance in CI repressor regulation of the bacteriophage 186 genetic switch. *Mol Microbiol.* 45, 697-710.

[11] Bunke, H. (1997). On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18, 689-694.

[12] Little, J.W. (1984) Autodigestion of lexA and phage lambda repressors. *Proc. Natl. Acad. Sci.* U S A. 81, 1375-1379.

[13] Kobiler, O., Rokney, A., Friedman, N., Court, D.L., Stavans, J., and Oppenheim, A.B. (2005). Quantitative kinetic analysis of the bacteriophage lambda genetic network. *Proc Natl Acad Sci U S A*, **102**, 4470-4475 (2005).

[14] Pedulla, M.L., Ford, M.E., Karthikeyan, T., Houtz, J.M., Hendrix, R.W., Hatfull, G.F., Poteete, A.R., Gilcrease, E.B., Winn-Stapley, D.A., and Casjens, S.R. (2003). Corrected sequence of the bacteriophage P22 genome. *J Bacteriol.* **185** (4): 1475-7.

[15] Kauffman, S.A. (1993). The Origins of Order: Self Organization and Selection in Evolution, Oxford University Press, New York.

[16] An example of this is the two paths from RecA to CII over respectively CI and LexA, of which only the RecA-LexA-CII path is activated during lysis.

[17] Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004) The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U S A* 101, 4781-4786.

[18] Hamoen, L.W., Venema, G., and Kuipers, O.P. (2003) Controlling competence in Bacillus subtilis: shared use of regulators. *Microbiology* 149, 9-17.

[19] Albert, I., and Albert R. (2004). Conserved network motifs allow protein-protein interaction prediction *Bioinformatics*, 20, 3346-3352.

[20] Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A. (1983). Lambda II. (eds.) sl Cold Spring Har-

bor Press, N.Y

[21] Kobiler, O., Oppenheim, A.B., Herman C. (2004) Recruitment of host ATP-dependent proteases by bacteriophage lambda. *J. Struct. Biol.* 146, 72-78.

[22] Metropolis, M., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys* 21, 1087-1092.

[23] Kirkpatrick, S., Gelatt, Jr.C.D., and Vecchi, M.P. (1983) Optimization by simulated annealing. *Science* 220, 671-680.